Holly Erickson DSC 540 Mid-Term

A 250-word paper summarizing your steps and any challenges you ran into during the project.

You should also outline any decisions you had to make while transforming the data (for

example, if you decided to remove duplicates, how did you arrive at that decision?).

Major Challenge 1: Finding the dataset.

I looked at quite a few while finding one that met the minimum requirements for number of features and records. The second criteria was for it to be something I’m interested in, as I find it helps make the analysis more fun. One interesting thing I found when looking at the Kaggle site is how much their competitions are moving towards image processing. I don’t remember this being the case when I first started doing Kaggle competitions, about a year ago. Anyways, I eventually found an awesome dataset on pizza restaurants.

Major Challenge 2: Using the library Agate-stats

I read in ch. 9 of “Data Wrangling with Python” that Agate-stats is a great tool to find outliers. I was able to install it using Anaconda, however it required Python version 3.6 or older, and I was currently using version 3.7. This seemed like a great opportunity to use the virtual environment feature of Anaconda. I have been curious about virtual environments, but never felt a need to use one prior to this. I created a new environment that used Python v. 3.6, and installed Spyder and Pandas (which auto-installs Numpy in Anaconda). From there I was able to install Agate-Stat, and continue on with the project using my “Agate-stats” virtual environment. Success!

Restaurants are listed on several rows, a row for each menu item we have data for. When transforming the data, I only printed restaurant details such as location once, and did not print duplicate information / save duplicate information to my readable text file.